W560–W568 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 17 May 2012*
doi:10.1093/nar/gks432

O-miner: an integrative platform for automated analysis and mining of -omics data

Rosalind J. Cutts, Abu Z. Dayem Ullah, Ajanthah Sangaralingam, Emanuela Gadaleta, Nicholas R. Lemoine and Claude Chelala*

Centre for Molecular Oncology, Barts Cancer Institute, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK

Received January 30, 2012; Revised April 19, 2012; Accepted April 24, 2012

ABSTRACT

High-throughput profiling has generated massive amounts of data across basic, clinical and translational research fields. However, open source comprehensive web tools for analysing data obtained from different platforms and technologies are still lacking. To fill this gap and the unmet computational needs of ongoing research projects, we developed O-miner, a rapid, comprehensive, efficient web tool that covers all the steps required for the analysis of both transcriptomic and genomic data starting from raw image files through in-depth bioinformatics analysis and annotation to biological knowledge extraction. O-miner was developed from a biologist end-user perspective. Hence, it is as simple to use as possible within the confines of the complexity of the data being analysed. It provides a strong analytical suite able to overlay and harness large, complicated, raw and heterogeneous sets of profiles with biological/clinical data. Biologists can use O-miner to analyse and integrate different types of data and annotations to build knowledge of relevant altered mechanisms and pathways in order to identify and prioritize novel targets for further biological validation. Here we describe the analytical workflows currently available using O-miner and present examples of use. O-miner is freely available at www.o-miner.org.

INTRODUCTION

High-throughput profiling platforms have produced a large amount of data with public repositories such as the Genome Expression Omnibus (1) and ArrayExpress (2,3) already storing tens of thousands of profiles across different experimental conditions. There is a steady growth in the amount and diversity of profiling results causing challenges in data analysis and integration as well as a strong need for novel comprehensive online bioinformatics tools which are easy to use by biologists and able to process raw profiles in a single- or global-analysis manner.

Although many methods are now available for low- and high-level analysis of genomic and transcriptomic experiments (4–7), most require programming knowledge as well as bioinformatics expertise and results can vary substantially amongst these. Analysis of large data sets may involve the need for powerful computational resources as well as time and effort to set up the necessary infrastructure. For example, the use of aroma.affymetrix (4,8) for analysis of copy number data involves the creation of annotation files via a specific directory with a strict directory structure to organize raw and processed data. Additionally, there is no analytical tool that can handle raw and/or partially processed genomics data and annotate/display results online in a user-friendly manner that would alleviate the need for bioinformatics expertise and allow researchers to process their in-house data in isolation or alongside the accumulated publicly available data in their area of research.

To overcome these problems, we have developed O-miner (http://www.o-miner.org), which can analyse the most popular, and widely used Affymetrix genomics and transcriptomics array types on the fly starting from raw standard Affymetrix file format (CEL image files obtained from the scanner) or partially processed format (normalized, segmented and/or binary) with minimal set-up efforts. The analysis is performed on a dedicated server removing memory or disk space requirements on end-user machines. All analytical pipelines are transparent, robust, well documented and based on well-established and recently developed statistical methods. Results can be viewed online as dynamic HTML reports for easy navigation through an interactive friendly interface or downloaded as text, excel or graphics files.

O-miner is comprehensive, robust, memory-efficient and can easily be extended with new methods and algorithms to cover additional chip types and platforms. In this article, we provide an overview of O-miner and discuss both transcriptomics and genomics workflows.

*To whom correspondence should be addressed. Tel: +44 207 882 3570; Fax: +44 207 882 3884; Email: c.chelala@qmul.ac.uk

We outline some examples of use to show how to perform low single-level as well as high global-level analysis and to illustrate how to navigate through obtained results. Finally, we discuss future updates of the software to accommodate and link additional data types.

OVERVIEW OF O-MINER

O-miner provides a framework for automated analysis of different types of -omics data and currently covers the analysis and annotation of both genomic and transcriptomic data. The user must first upload the data files to be analysed as a zip archive to the O-miner server using the graphical interface or enter a valid GEO series number (GSE format). This alleviates the time-consuming and repetitive task of uploading one data file at a time. Once data transfer is completed, the 'File Organizer' window displays the individual files and can be used for the assignment of sample names and biological groups before specifying the analysis options. A unique project is created for each submitted analysis.

Genomics analysis

A general copy number analysis pipeline starts from probe level raw intensity .CEL data files obtained immediately after scanning, through background adjustment, normalization and summarization to derive raw copy number data (normalized log2 ratio sample/reference format) followed by segmentation and smoothing (segmented format) before thresholding and calling regions of copy number gain and loss (binary format). To the best of our knowledge, O-miner is the only freely available web tool that can accept data submission at any stage of this pipeline either as .CEL files or partially processed (normalized, segmented or binary) data files.

Raw CEL intensity files

O-miner enables the two common scenarios of copy number analysis. The first is a paired analysis option where each sample is coupled with a specific unique reference (e.g. a cancer sample with its corresponding matched normal sample). The second is an unpaired analysis where each sample uses the same common reference, which is often the average of a pool of samples. Both options are possible on a wide variety of Affymetrix platforms including the widely used GeneChip® Human Mapping Arrays 10K, 100K Set (50K_Hind240 and 50K_Xba240) and 500K Set (250K_Nsp and 250K_Sty) as well as Genome-Wide Human SNP Arrays 5.0 and 6.0. We have processed and made available precompiled raw HapMap data (CEL files) from four human populations: African YRI (from Yoruba in Ibadan, Nigeria), Japanese JPT (from Tokyo, Japan), Han Chinese CHB (from Beijing, China) and European CEU (from Utah, USA with ancestry from northern and western Europe) to use as a baseline in an unpaired analysis scenario. After extracting the zip archive, O-miner displays the available .CEL files list in the 'File Organizer' for the user to create Sample/Reference attributes and define Samples and References lists. The first file in the Sample Files list is compared to the first file in the Reference Files list, the second files in both lists are compared to each other and so on. In the same manner, files from different enzyme sets for Human Mapping 100K/500K array sets can be paired to match and merge array sets originating with the same sample. The Sample/Reference attributes are not required for unpaired analysis. Data are categorized by entering a biological group attribute to define the biological source/state at the origin of each array (primary, metastasis, resistant, etc.). O-miner combines the results observed in the same biological source/state and performs group comparisons.

O-miner reads CEL intensity files and automatically builds up the required directory structure and annotation files to run the methods implemented in the aroma.affymetrix framework (4,8). Briefly, O-miner performs initial quality control checks, background correction, allelic cross-talk calibration, nucleotide-position probe sequence effects normalization, probe-level summarization using robust average (for SNP 5.0 and 6.0 arrays) or log-additive model (for 10, 100 and 500K arrays), PCR fragment-length effects normalization and calculates raw copy number estimates (log2 ratios) relative to the chosen reference. These normalized estimates are used as input for segmentation methods to identify copy number regions and further subsequent analysis as explained below.

Partially processed (normalized, segmented or binary)

Normalized data text files as obtained from the raw CEL analysis described above or from other normalization methods and algorithms can be used. If uploaded as a new submission, the 'File Organizer' extracts the sample names from the column headings of the uploaded file and offers the option to enter a biological group attribute to define the biological source/state at the origin of each sample for further subgroup analysis. At this level, O-miner is ready to apply a segmentation analysis by offering 10 popular algorithms as implemented in the R package CGHweb (5). Briefly, these are BioHMM, CBS, FASeg, cghFLasso, CGHseg, GLAD, LOWESS, Wavelet smoothing, Quantile Smoothing and Running Average (9–17). The user selects the method(s) to be used to derive a consensus profile from multiple probes/samples. Added to the benefit of assessing segmented profiles from different algorithms, this also offers the user the possibility of checking whether a copy number alteration arose as an artefact of the specified segmentation method. The results are then ready to be processed to determine the regions of gains/losses according to user-defined cut-offs based on the log2 ratio threshold value, consecutive number of SNPs that form a copy number region (at least 15 SNPs by default) and frequency of samples where a copy number event was observed (at least 20% by default). O-miner offers an option to predict the log2 ratio threshold based on the quantile distribution of segmented raw copy numbers. Once a threshold is determined the data could be binary coded (0: no changes, 1: copy number gain, −1: copy number loss) for subsequent analysis. Similarly, users can start their data analysis from this level by submitting a binary coded data file.

Further analysis options

Once regions of gains/losses have been determined, O-miner can provide physical and cytogenetic mapping information as well as related gene annotations from UCSC (18), NCBI RefSeq (19), Ensembl (20) and VEGA (21). O-miner also investigates regulatory elements, such as conserved Transcription Factor Binding Sites (22) and microRNA (23,24). As disease/critical genes are more likely to be located in copy number regions that are common/recurrent among samples, O-miner provides the analysis option of identifying recurrent regions of copy number alterations within the biological groups being investigated. These minimum common regions (MCR) can be calculated by using one of the three robust methods: CGHregions (25), RJaCGH (26) and MSA (27).

Transcriptomics analysis

Expression profiling analysis starts from probe level raw intensity .CEL data files obtained immediately after scanning, through background correction, normalization and summarization to derive expression measurements data (normalized data matrix) followed by filtering to reduce data dimensionality and differential analysis to detect de-regulated genes. O-miner accepts data submission as .CEL files, normalized or filtered data matrix files. O-miner enables the analysis of paired samples/replicates.

Raw CEL intensity files

A wide variety of Affymetrix platforms including the widely used GeneChip® Human Genome Arrays U95 Set (U95A, U95Av2, U95B, U95C, U95D, U95E), U133 Set (U133A and U133B), U133A 2.0 and U133 Plus 2.0 are available. After extracting the uploaded archive of array files, O-miner displays the available .CEL files list in the 'File Organizer' window for the user to define the samples list and biological source/state at the origin of each array. If performing a paired analysis, the user needs to arrange the samples in pairs in the related two group lists in the 'File Organizer'. If the experiment contains technical replicates, the user must indicate the replicates in the additional 'Replicate' column that will appear in the 'File Organizer'. O-miner combines the results observed in the same biological source/state and performs differential analysis between selected groups.

O-miner reads CEL intensity files and runs the quality control (QC) methods implemented in the R package ArrayMvout (28) to automatically exclude outliers from subsequent analysis. An additional manual check could be performed using ArrayQualityMetrics (29). This is followed by normalization using RMA (30), GCRMA (31) or tRMA (32). These normalized estimates are used as input for filtering and differential analysis methods to identify de-regulated expression and run further analyses as outlined below.

Normalized or filtered data

Normalized data text files as obtained from the raw CEL analysis described above or from other normalization methods can be used. If uploaded as a new submission, the 'File Organizer' window displays the sample names as extracted from the uploaded file and offers the option to enter a biological group attribute to define the biological source/state at the origin of each sample for further subgroup analysis. At this level, O-miner is ready to apply a filtering step to reduce the dimensionality of the data by offering three popular methods: interquartile range (IQR) (soft, intermediate, robust), intensity (25% or 50% of samples above 100) or standard deviation (top 10% or 5% most variable probes). Differential expression analysis is applied to the filtered matrix using LIMMA (33). O-miner will automatically refresh to display a 'LIMMA comparison' section with the list of biological groups allowing the user to define the contrast and design matrices required by LIMMA based on the user selection of the comparisons between the predefined biological groups. A number of statistics for differential expression are provided to refine the de-regulated genes list according to user-defined cut-offs based on log2 fold change values (2 by default) and *P*-values (0.05 by default) adjusted using Holm (34), Benjamini and Hochberg (BH) (also known as FDR) (35) or Benjamini and Yekutieli (BY) (36) multiple testing correction methods.

Further analysis options

GOstats (37) can be used to assess the overrepresentation of GO terms among the GO annotations for the differentially expressed genes. Additional expression plots can also be generated from the results page allowing the user to examine the level/change in expression among the experimental datasets for a particular gene(s)/probe(s) of interest in the filtered data. A Venn diagram for up to four biological groups can be produced to show the common and specific differentially expressed probes (all, up- or down-regulated).

EXAMPLES OF USE

O-miner provides comprehensive interactive web pages in a tabbed browsing format that are intended to guide the user through the key results for their analysis. All data are also available to download and view locally as text, excel or image files.

Genomics

Results are displayed as a tabbed view representing QC, clustering, MCR (if selected), sample and group information. Using the 'Sample View' it is possible to browse through the results obtained for each individual sample including log2ratio plots and annotated regions of gains and losses that can also be viewed as a track in the UCSC Genome Browser alongside a rich collection of annotations. The 'Group View' summarizes results based on the biological groups originally defined by the user including frequency plots and a gene-level view to summarize the gene content within copy number alterations.

To demonstrate the functionality of O-miner, we analysed 25 samples from mutated (KIT or PDGFRA) or wild-type gastrointestinal stromal tumours (GISTs) profiled using Affymetrix Genome-Wide Human SNP 6.0 platform (GSE20709). We applied an unpaired analysis using the wild-type patients as baseline. We

used Picard, Fused Lasso and CBS algorithms for segmentation and applied a minimum physical length of at least 15 consecutive SNPs for putative regions of genetic alterations. The threshold for gains or losses was determined by O-miner based on the inspection of the quantile distribution of the segmented ratios. O-miner provides straightforward access to results for each biological group, and an easy way to drill down to individual results for a specified sample. The 'Sample View' and 'Group View' of obtained results with related mining options are presented in Figures 1 and 2, respectively. One can navigate through putative regions of gains and losses, frequency plots for a specified sample and automatically view this information within the UCSC Genome Browser where we could zoom in to specific regions of interest. This allows the data to be mined and visualized alongside a large collection of annotation data tracks. Our results clearly show the hot spots for copy number loss on chromosomes 1, 14 and 22 as previously reported. The 'Group View' can easily be used to overlay and compare results from the two mutated sample sets (Figure 3).

To demonstrate further capabilities of O-miner, we analysed a panel of 12 primary effusion lymphoma (PEL) cell lines profiled with the Affymetrix GeneChip® Human Mapping Arrays 500K Set (GSE28684) (38) using an unpaired analysis with normal tonsil controls. Segmentation and thresholding methods were defined as in the previous example. We replicated the previously reported PEL-associated genomic amplifications in chromosome 1q, 7, 8 and 12. Furthermore, as the majority of PEL are co-infected with Epstein–Barr virus (EBV), we segregated PEL samples into EBV-positive and EBV-negative subgroups and investigated the recurrent copy number alterations in each group using MSA. As shown in Figure 4A, one could quickly compare and visualize MCR plots at the genome or chromosome level for both biological groups. Results are available in HTML, Excel or Bed formats. Results can also be viewed in the UCSC Genome Browser (Figure 4B), where we compared a detected MCR region on chromosome 19p13.3 across the biological subgroups and investigated its gene content. In a few seconds, this visual inspection narrowed down an MCR of genetic gain specific to the EBV-negative subgroup. By displaying the RefSeq annotation track, we directly pointed to RFX2, ACSBG2 and FUT3 genes reported in the original study to be altered only in the EBV-negative PEL subgroup. We also identified few other important genes mapping to this MCR and also relevant to EBV-negative PEL subgroup.

In addition to analysing data from individual studies, O-miner provides a high-level analysis option. Data from multiple sources/formats can be merged at different levels (.CEL files, normalized, segmented or binary) and submitted to O-miner. This gives the user increased flexibility for carrying out a global analysis dependent on the data types available. For example if .CEL files are not available, it is possible to submit merged partially analysed data (normalized, segmented or binary coded format). This also provides a method of submitting larger datasets.

Transcriptomics

Results are displayed as a tabbed view representing QC, clustering, differential expression, gene ontology

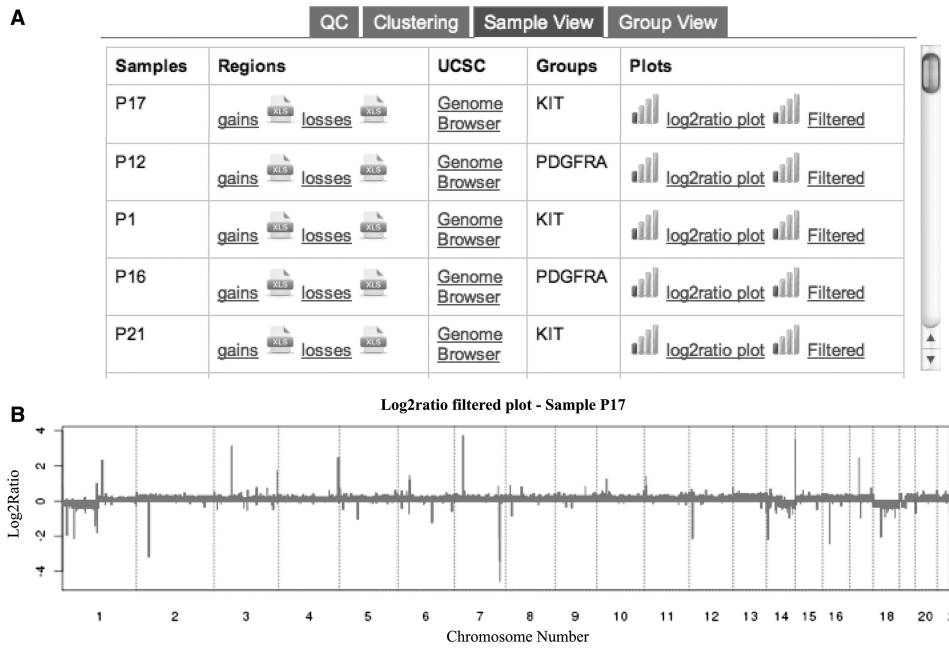

Figure 1. 'Sample View' of the results produced by O-miner for the GIST data (GSE20709). (**A**) From left to right, each column present the following information: 'Samples' display the identifiers as given by user, 'Regions' provides links to display annotated regions of gains and losses in HTML and Excel formats, 'UCSC' displays the results as a track in the UCSC Genome Browser alongside a rich collection of public annotations, 'Groups' represent biological groups as defined by user and 'Plot' provides links to produce log2ratio plot for each sample with or without filtering (HTML or image file). For example, clicking on Filtered for sample P17 will produce the log2ratio plot of filtered data in the same page (**B**).

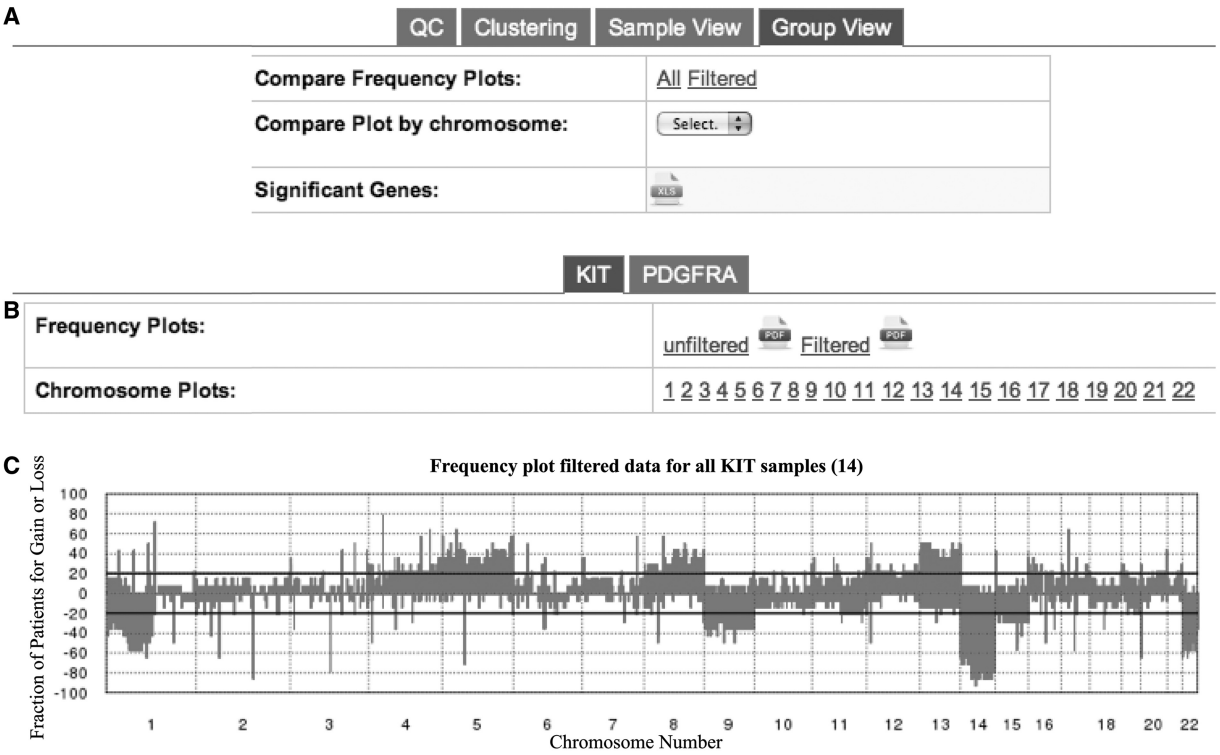

Figure 2. 'Group View' of the results produced by O-miner for the GIST data (GSE20709). (**A**) Users can compare frequency plots between the different biological groups with or without filtering, for the whole genome or a specific chromosome and obtain a list of the significant genes within the frequently altered copy number regions. (**B**) Another tabbed view enables users to browse between the defined biological groups and view frequency plots for each biological group with or without filtering, at the genome or chromosome level in HTML or pdf formats. For example, clicking on Filtered for KIT subgroup will produce the log2ratio plot of filtered data for the 14 KIT mutated samples in the same page (**C**).

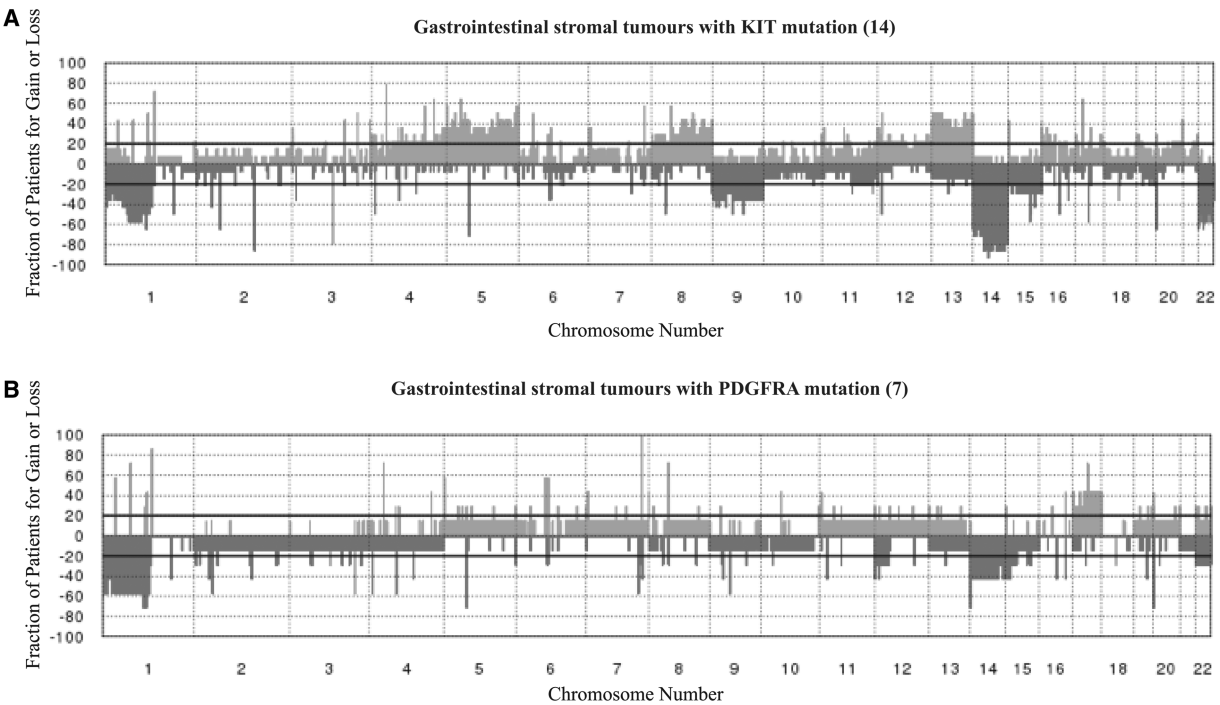

Figure 3. 'Group View', Compare Frequency Plots option, for the biological groups within the GIST data (GSE20709). O-miner provides a useful summary of putative regions of copy number gains and losses by providing frequency plots for each defined biological group. The user can easily compare and contrast results from the two biological groups in this study [14 samples with KIT mutation (**A**) and 7 samples with PDGFRA mutation (**B**)].

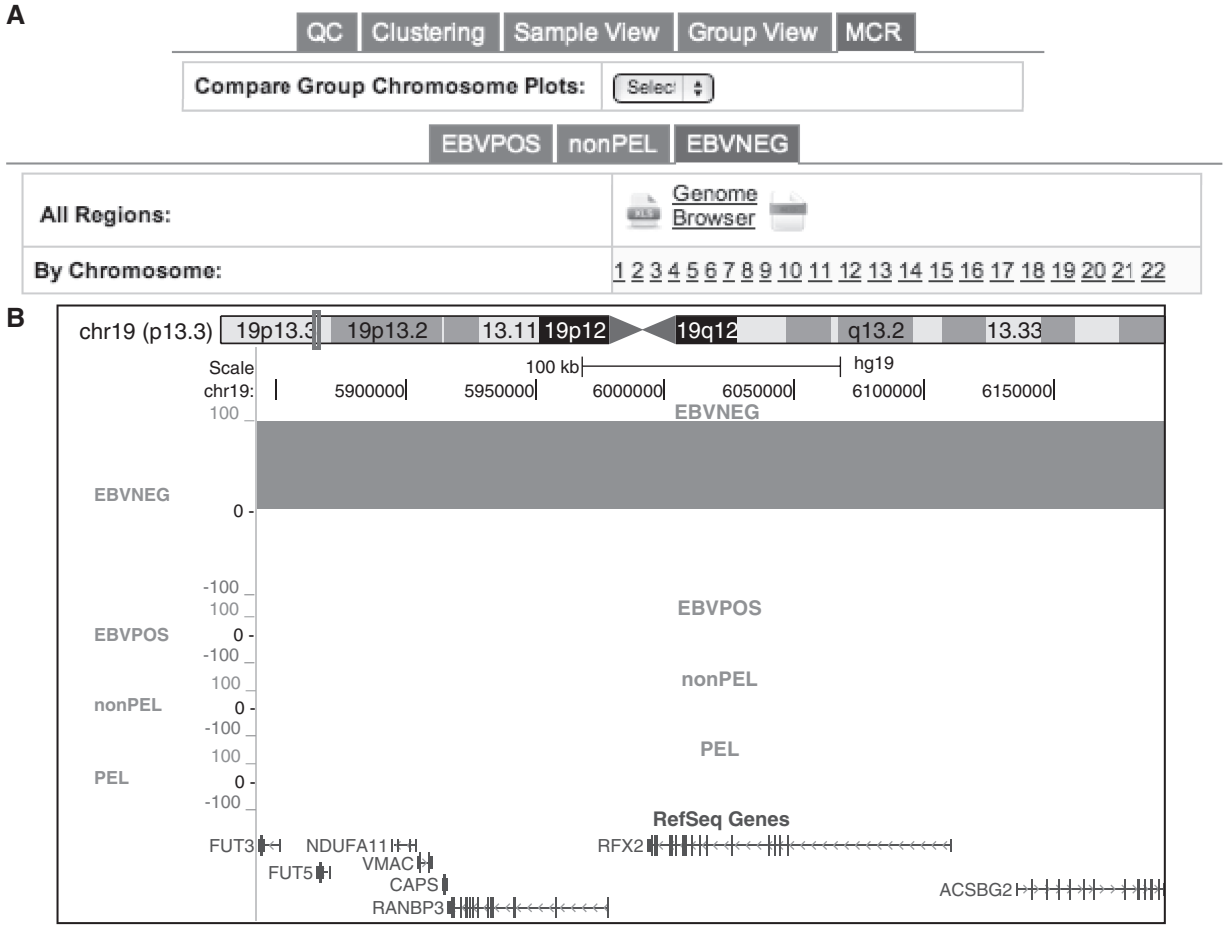

Figure 4. 'MCR View' of PEL (EBV-negative and EBV-positive) and non-PEL cell lines (GSE28684). (**A**) It is possible to identify recurrent regions of copy number alterations within the biological groups being investigated, compare chromosome plots and explore MCR regions for each biological group in more detail for the whole genome or a specific chromosome. Results could be exported as Excel or BED files. Results could also be viewed in the UCSC Genome Browser (**B**), where one could overlay and compare the detected MCR regions in each biological group, zoom in a specific MCR region on chromosome 19p13.3 and investigate its gene content using the RefSeq genes track. In a few seconds, a quick visual inspection narrowed down a smaller region on 19p13.3 where there is an MCR of genetic gain specific to the EBV-negative PEL subgroup. This directly points to some of the genes reported in the original paper (RFX2, ACSBG2 and FUT3) as well as others important ones mapping to this MCR.

(if selected) and expression plots. As an example, we analysed six drug-resistant/parental MIA-PaCa-2 pancreatic cell lines profiled using Affymetrix GeneChip® Human Genome Arrays U133 Plus 2.0 (GSE16648) (39). After applying QC, normalization using GCRMA and filtering by standard deviation to select the top 5% of most variable probes, we performed a differential expression analysis using LIMMA to compare resistant to parental cell lines. A typical O-miner tabbed output includes QC information, differentially expressed genes, a cluster dendrogram, overrepresented gene ontology terms and an expression plot generator that could be used to produce expression plots on the fly to compare the expression level of a gene(s)/probe(s) of interest across the array data within the defined biological groups (Figure 5).

O-miner can also be used to run a rapid global analysis on transcriptomics data. For example, we analysed .CEL files from three prostate cell lines (LNCaP, DU145 and PC3) from three different studies in ArrayExpress/GEO (E-TABM-948, GSE32474 and E-GEOD-28846). Figure 6 demonstrates additional O-miner output capabilities and shows a Venn diagram indicating the overlap of differentially expressed probes between the different cell lines and clustering of expression data across the experimental groups.

CONCLUSIONS AND FUTURE WORK

O-miner is a useful and flexible tool, particularly for biologists to carry out routine data analysis without the need for a complex IT infrastructure or in-depth bioinformatics support. Future plans include the addition of further analysis pipelines, in particular for methylation, miRNA and downstream mining of next-generation sequencing data. In its current version, O-miner allows users to submit data by giving the GEO series number. For the moment this is limited to series with samples profiled on

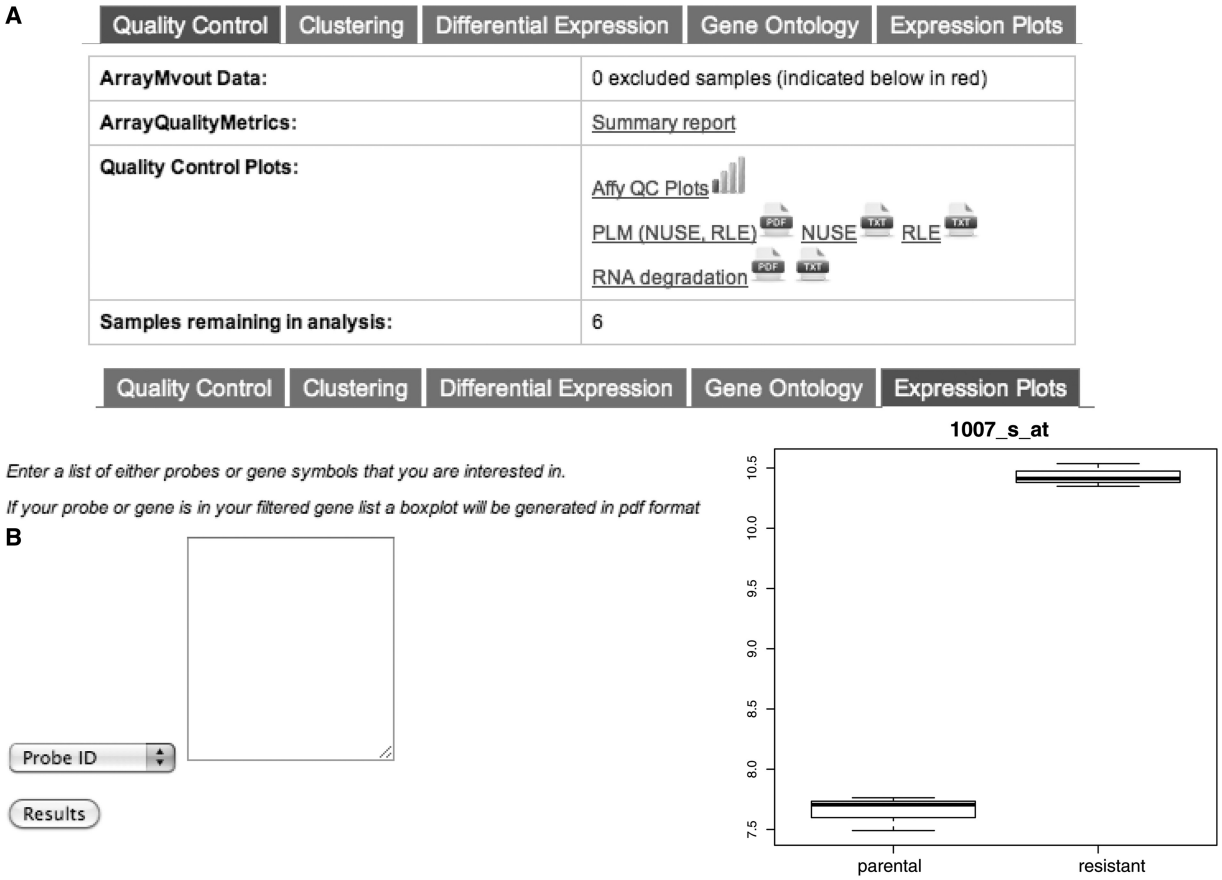

Figure 5. Expression analysis of MIA-PaCa-2 resistant/parental samples (GSE16648). O-miner produces a tabbed view of results with quality control (**A**), clustering, differential expression, gene ontology and expression plot generator for a particular gene(s)/probe(s) of interest (**B**).

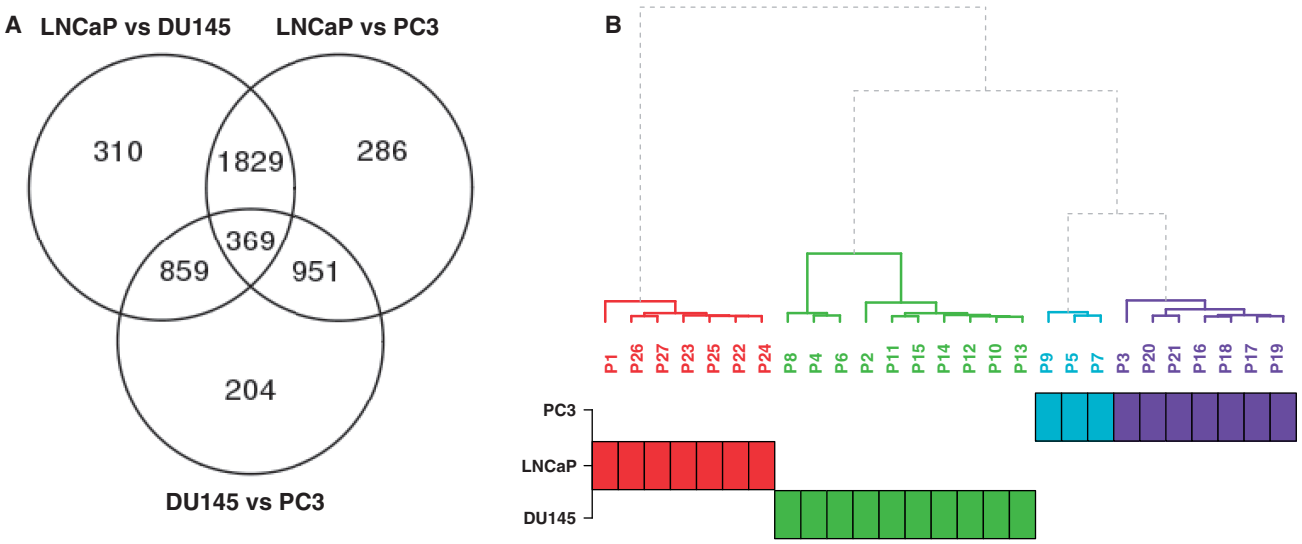

Figure 6. Global-analysis of prostate cell lines from three studies (E-TABM-948, GSE32474 and E-GEOD-28846). (**A**) Venn diagram showing overlaps between differentially expressed probes in each comparison. (**B**) Coloured cluster dendrogram. Each cluster has its colour. The plot displays the biological groups below in order to quickly compare it with the observed clusters.

the same platform. We plan to develop this further in future releases. We are also planning to cover additional platforms/species such as Illumina and Affymetrix Whole-Transcript arrays and to make O-miner available as an R package.

ACKNOWLEDGEMENTS

Authors thank their colleagues who have tested O-miner.

FUNDING

Breast Cancer Campaign (to R.J.C.); Cancer Research UK (to A.S and A.Z.D.U). Funding for open access charge: Cancer Research UK [programme grant reference 15310].

Conflict of interest statement. None declared.

REFERENCES

1. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al*. (2010) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res*., **39**, D1005–D1010.
2. Parkinson,H., Kapushesky,M., Kolesnikov,N., Rustici,G., Shojatalab,M., Abeygunawardena,N., Berube,H., Dylag,M., Emam,I., Farne,A. *et al*. (2009) ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*., **37**, D868–D872.
3. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Holloway,E. *et al*. (2011) ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*., **39**, D1002–D1004.
4. Bengtsson,H., Irizarry,R., Carvalho,B. and Speed,T.P. (2008) Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, **24**, 759–767.
5. Lai,W., Choudhary,V. and Park,P.J. (2008) CGHweb: a tool for comparing DNA copy number segmentations from multiple algorithms. *Bioinformatics*, **24**, 1014–1015.
6. Carro,A., Rico,D., Rueda,O.M., Diaz-Uriarte,R. and Pisano,D.G. (2010) waviCGH: a web application for the analysis and visualization of genomic copy number alterations. *Nucleic Acids Res*., **38**, W182–W187.
7. Medina,I., Carbonell,J., Pulido,L., Madeira,S.C., Goetz,S., Conesa,A., Tarraga,J., Pascual-Montano,A., Nogales-Cadenas,R., Santoyo,J. *et al*. (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res*., **38**, W210–W213.
8. Bengtsson,H., Simpson,K., Bullard,J. and Hansen,K. (2008) aroma.affymetrix: a generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. *Tech Report #745*. Department of Statistics, University of California, Berkeley, February 2008.
9. Eilers,P.H. and de Menezes,R.X. (2005) Quantile smoothing of array CGH data. *Bioinformatics*, **21**, 1146–1153.
10. Hsu,L., Self,S.G., Grove,D., Randolph,T., Wang,K., Delrow,J.J., Loo,L. and Porter,P. (2005) Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**, 211–226.
11. Hupe,P., Stransky,N., Thiery,J.P., Radvanyi,F. and Barillot,E. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.
12. Jones,L., Goldstein,D.R., Hughes,G., Strand,A.D., Collin,F., Dunnett,S.B., Kooperberg,C., Aragaki,A., Olson,J.M., Augood,S.J. *et al*. (2006) Assessment of the relationship between pre-chip and post-chip quality measures for Affymetrix GeneChip expression data. *BMC Bioinformatics*, **7**, 211.
13. Marioni,J.C., Thorne,N.P. and Tavare,S. (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, **22**, 1144–1146.
14. Olshen,A.B., Venkatraman,E.S., Lucito,R. and Wigler,M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.
15. Picard,F., Robin,S., Lavielle,M., Vaisse,C. and Daudin,J.J. (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, 27.
16. Tibshirani,R. and Wang,P. (2008) Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, **9**, 18–29.
17. Yu,T., Ye,H., Sun,W., Li,K.C., Chen,Z., Jacobs,S., Bailey,D.K., Wong,D.T. and Zhou,X. (2007) A forward-backward fragment assembling algorithm for the identification of genomic amplification and deletion breakpoints using high-density single nucleotide polymorphism (SNP) array. *BMC Bioinformatics*, **8**, 145.
18. Hsu,F., Kent,W.J., Clawson,H., Kuhn,R.M., Diekhans,M. and Haussler,D. (2006) The UCSC Known Genes. *Bioinformatics*, **22**, 1036–1046.
19. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **35**, D61–D65.
20. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al*. (2012) Ensembl 2012. *Nucleic Acids Res*., **40**, D84–D90.
21. Wilming,L.G., Gilbert,J.G., Howe,K., Trevanion,S., Hubbard,T. and Harrow,J.L. (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res*., **36**, D753–D760.
22. Wingender,E. (2008) The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform*., **9**, 326–332.
23. Grimson,A., Farh,K.K., Johnston,W.K., Garrett-Engele,P., Lim,L.P. and Bartel,D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*., **27**, 91–105.
24. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res*., **36**, D154–D158.
25. van de Wiel,M.A. and Wieringen,W.N. (2007) CGHregions: dimension reduction for array CGH data with minimal information loss. *Cancer Inform*., **3**, 55–63.
26. Rueda,O.M. and Diaz-Uriarte,R. (2009) RJaCGH: Bayesian analysis of aCGH arrays for detecting copy number changes and recurrent regions. *Bioinformatics*, **25**, 1959–1960.
27. Guttman,M., Mies,C., Dudycz-Sulicz,K., Diskin,S.J., Baldwin,D.A., Stoeckert,C.J. Jr and Grant,G.R. (2007) Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genet*., **3**, e143.
28. Asare,A.L., Gao,Z., Carey,V.J., Wang,R. and Seyfert-Margolis,V. (2009) Power enhancement via multivariate outlier testing with gene expression arrays. *Bioinformatics*, **25**, 48–53.
29. Kauffmann,A., Gentleman,R. and Huber,W. (2009) arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics*, **25**, 415–416.
30. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
31. Wu,Z., Irizarry,R., Gentleman,R., Martinez Murillo,F. and Spencer,F. (2004) A model based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc*., **99**, 909–917.
32. Giorgi,F.M., Bolger,A.M., Lohse,M. and Usadel,B. (2010) Algorithm-driven artifacts in median Polish summarization of microarray data. *BMC Bioinformatics*, **11**, 553.
33. Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*., **3**, Article3.

34. Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.
35. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.
36. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
37. Falcon,S. and Gentleman,R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.
38. Roy,D., Sin,S.H., Damania,B. and Dittmer,D.P. (2011) Tumor suppressor genes FHIT and WWOX are deleted in primary effusion lymphoma (PEL) cell lines. *Blood*, **118**, e32–39.
39. Selga,E., Oleaga,C., Ramirez,S., de Almagro,M.C., Noe,V. and Ciudad,C.J. (2009) Networking of differentially expressed genes in human cancer cells resistant to methotrexate. *Genome Med.*, **1**, 83.